DOCUMENT RESUME

ED 334 230 TM 016 738

AUTHOR Donoghue, John R.; Allen, Nancy L.
TITLE "Thin" versus "Thick" Matching in the Mantel-Haenszel Procedure for Detecting DIF.
PUB DATE Apr 91
NOTE 36p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 4-6, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; Difficulty Level; *Equations (Mathematics); *Item Bias; Item Response Theory; *Mathematical Models; Monte Carlo Methods; Sample Size; Test Format
IDENTIFIERS Item Discrimination (Tests); *Mantel Haenszel Procedure

ABSTRACT

This Monte Carlo study examined strategies for forming the matching variable for the Mantel-Haenszel (MH) differential item functioning (DIF) procedure. Data were generated using a three-parameter logistic item response theory model, with common guessing parameters. The number of subjects and test length were manipulated, as were the difficulty, discrimination, and presence of DIF in the studied item. Each replication of a test represented a test administration. Replications were made at two levels: replication within test administrations, and replication of responses to the studied items for each test administration. There were 20 replications for responses to the studied item, and 20 replications of administrations within test conditions. "Thin matching" (the use of the total score as the matching variable for MH) was compared to forms of "thick matching" (forming the matching variable by pooling total score levels). Outcome measures included the transformed log-odds delta(sub MH), standard error (delta(sub MH), and the MH chi square. The results indicate that "thick matching" can improve the performance of the MH procedure for detecting DIF. For short tests (5 to 10 items), "thin matching" yielded very poor results, with a tendency to falsely identify items as possessing DIF against the reference group. For long tests (40 items), especially with adequate sample sizes (1,600), "thin matching" yielded the best results of any method examined. Tests of intermediate lengths yielded similar results for "thin matching" and the best methods of "thick matching". Implications for the use of auxiliary information in DIF studies are discussed. Ten data tables and a 21-item list of references are included. (RLC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED334230

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

JOHN R. DONOGHUE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

"Thin" versus "Thick" Matching in the Mantel-Haenszel Procedure for Detecting DIF

John R. Donoghue and Nancy L. Allen

Educational Testing Service

Paper presented at the Annual Meeting
of the National Council on Measurement in Education,
Chicago, IL, April 1991.

TM016738

BEST COPY AVAILABLE

ABSTRACT

This Monte Carlo study examined strategies for forming the matching variable for the Mantel-Haenszel DIF procedure. Data were generated with a 3PL IRT model, with common guessing parameter. Number of subjects and test length were manipulated, as were the difficulty, discrimination, and presence of DIF in the studied item. "Thin matching" on total test score was compared to forms of "thick matching," pooling levels of the total score. Outcome measures were the transformed log-odds $\hat{\Delta}_{MH}$, se($\hat{\Delta}_{MH}$) and the MH χ^2. Implications for the use of auxiliary information in DIF studies are discussed.

The study of differential item functioning (DIF), or item bias, is an important area in applied testing. DIF studies are typically concerned with comparing the relative performance of the subgroup of interest (the focal group) to that of a comparison or reference group. DIF occurs when members of the two groups possessing the same proficiency have different probabilities of getting the studied item correct (e.g., Lord, 1980). In assessing DIF, it is important to separate differences in proficiencies of the subgroups from the functioning of the item. Simply comparing the probability of a correct response (p-values) of the two groups usually is not a satisfactory procedure for assessing DIF. This comparison would assess the overall impact of the item, and is only valid as a measure of DIF if the two groups have identical proficiency distributions. When the two groups differ in proficiency, comparison of p-values confounds DIF with impact. To accurately assess DIF, it is necessary to condition on some measure of proficiency.

The MH DIF procedure

The Mantel-Haenszel (MH) statistic (1959) has gained widespread use as a measure of DIF since Holland (1985) and Holland and Thayer (1988) proposed its use for that purpose. Holland and Thayer (1988) and Thissen, Steinberg and Wainer (1988) examined the relationship between the MH and Item Response Theory (IRT) based procedures for assessing DIF, and concluded that the MH procedure compared favorably to the more complex and expensive IRT based procedure for detecting uniform DIF.

The MH procedure functions by matching the reference and focal groups on some measure of performance, termed the "matching variable." In usual DIF applications of MH, the matching variable is the total score on the test. For each level of the K levels of the matching variable, MH forms a 2 X 2 table:

Group	Performance on Studied Item		Total
	Passed Item	Failed Item	
Reference	A_k	B_k	n_{Rk}
Focal	C_k	D_k	n_{Fk}
Total	m_{1k}	m_{0k}	T_k

T_k is the total number of subjects at level k of the matching variable, n_{Rk} and n_{Fk} are, respectively, the number of reference and focal group members at level k, and m_{1k} is the number of examinees at level k of the matching variable who got the studied item correct, while m_{0k} is the number who missed the item.

The MH statistic $\hat{\alpha}_{MH}$ estimates α_{MH}, the pooled odds-ratio, and is computed:

$$\hat{\alpha}_{MH} = \frac{\sum_{k=1}^{K} A_k D_k / T_k}{\sum_{k=1}^{K} B_k C_k / T_k} \quad . \tag{1}$$

The MH statistic is often transformed in psychometric applications to $\hat{\Delta}_{MH}$:

$$\hat{\Delta}_{MH} = -2.35 * \log_e(\hat{\alpha}_{MH}) \tag{2}$$

This transformation makes the measure negative for items which are harder for the focal group, and puts $\hat{\Delta}_{MH}$ on the "delta-scale" used at ETS to measure item difficulty. The standard error of $\log_e(\hat{\alpha}_{MH})$ was studied by Phillips and Holland (1987). The standard error of $\hat{\Delta}_{MH}$ is given by Holland and Thayer (1988) as:

$$se(\hat{\Delta}_{MH}) = 2.35 * \sqrt{Var(\log_e(\hat{\alpha}_{MH}))} \tag{3}$$

where:

$$Var(\log_e(\hat{\alpha}_{MH})) = \frac{\sum_{k=1}^{K} \frac{U_k V_k}{T_k^2}}{2\left(\sum_{k=1}^{K} \frac{A_k D_k}{T_k}\right)^2} \tag{4}$$

$$U_k = (A_k D_k) + \hat{\alpha}_{MH}(B_k C_k)$$

$$V_k = (A_k + D_k) + \hat{\alpha}_{MH}(B_k + C_k)$$

A one degree of freedom MH χ^2 measure of the null hypothesis of no difference between the performance of comparable members of the focal and reference groups (Holland, 1985; Holland and Thayer, 1988) is also available:

$$\chi^2_{MH} = \frac{\left(\left|\sum_{k=1}^{K} A_k - \sum_{j=1}^{K} E(A_j)\right| - \frac{1}{2}\right)^2}{\sum_{j=1}^{K} Var(A_j)} \tag{5}$$

where:

$$Var(A_k) = \frac{n_{Rk} n_{Fk} m_{1k} m_{0k}}{T_k^2 (T_k - 1)} \tag{6}$$

$$E(A_k) = \frac{n_{Rk} m_{1k}}{T_k}$$

Using $\hat{\Delta}_{MH}$ divided by its standard error, or the MH χ^2 statistic, a decision rule may be developed to classify items as possessing DIF or not.

Holland and Thayer (1988) also demonstrated, for data which follow a Rasch model, the counter-intuitive result that it is necessary to include the studied item in the matching variable in order to derive an accurate measure of DIF. Zwick (1990) examined the case of data which follow a very general IRT model, and reached the same conclusion. Recently, a Monte Carlo study by Donoghue, Holland and Thayer (1989) examined the effect of a variety of "nuisance" test characteristics on the functioning of the MH procedure. The study also confirmed the appropriateness of including the studied item in the matching variable. The score for items, other than the studied item, which contribute to the matching variable, is called the criterion.

"Thin" versus "Thick" Matching

The use of total score as the matching variable ("thin matching") for MH is a natural choice, but it is also somewhat arbitrary. Forming the matching variable by pooling total score levels ("thick matching") has been suggested by several authors. Other χ^2 based procedures for assessing DIF pool total scores in differing ways (Camilli 1979, in Crocker and Algina, 1986; Scheuneman, 1978, 1981). These procedures use fewer levels of the matching variable (typically 4-10), and compute a multiple degree of freedom χ^2.

The use of such "thick" matching has three potential advantages. First, estimation of the cell frequencies for individual levels of the matching variable becomes more stable and precise. This may be less of an advantage for the MH χ^2 than for the other χ^2 indices, as the MH procedure incorporates a weighting function which assigns less weight to those levels of the matching variable which have cells with small frequencies.

The second potential advantage of thick matching is related to the first. Thick matching allows all, or at least more of the data to be used. The MH weighting function assigns zero weights to those levels of the matching variable for which at least one of the cells has a zero frequency, effectively throwing away those observations. Thus, overly fine matching can result in elimination of much of the data. Further, if the test is very hard or very easy for one of the groups, thin matching on total score may result in a large number of levels being assigned zero weights, with a concomitant loss of data, and the MH statistic being based on

relatively few levels of the matching variable. An alternative to losing this data is to pool extreme levels of the matching variable until all levels have positive weights, as was done by Zwick and Ercikan (1989) in their study of NAEP history items.

The third potential advantage of thick matching is increased reliability of the matching variable. The reasoning is that assignment to score intervals is more reliable than assignments to the individual scores. Hence, matching on the more reliable intervals is likely to yield a more valid test for DIF.

The chief disadvantage of thick matching is the confounding of DIF with impact. In the extreme, there is only one level of the matching variable and the MH procedure using thick matching compares the simple p-values for each group. In this case, the confounding of DIF and impact is maximal. As the number of levels of the matching variable increases, the MH procedure differs more and more from an analysis of impact.

The effect of thick matching on MH statistics is not known. Some evidence of its characteristics may be gleaned by considering the behavior of the MH procedure for sparse 2 X 2 X K tables. As stated earlier, using thick matching to increase the stability of individual cell frequencies may be less of an advantage for the MH χ^2 than for the other χ^2-based indices, as the MH procedure contains an implicit weighting function which assigns less weight to those levels of the matching variable which have cells with small frequencies. Still, Camilli and Smith (1990) express concern over the effectiveness of the MH procedure for sparse tables in the presence of disparate numbers of subjects and relatively large differences in the proficiency distributions of the focal and reference groups. Camilli and Smith (1990) did find that the MH χ^2 compared favorably to both randomization and jackknife based estimators for sparse tables. Also, Breslow (1981) has shown that the efficiency of the estimator of the common odds-ratio increases dramatically as the number of sparse 2 X 2 tables increases and that $\hat{\alpha}_{MH}$ is asymptotically efficient for sparse 2 X 2 tables. On the other hand, Donner and Hauck (1986) have shown that efficiency increases as the sample size within a fixed number of 2 X 2 tables increases.

Strategies for Pooling Scores

Several strategies for forming the matching variable have been suggested, some in the context of other χ^2-based procedures for assessing DIF (e.g., Camilli (1979) as discussed in Crocker and Algina, 1986; Scheuneman, 1979, 1981). We will consider seven major methods in this paper. In the first strategy, thin matching, the matching variable is the simple total score. This is the strategy typically used for Mantel-Haenszel DIF studies.

The second strategy for thick matching, equal interval matching, involves simply dividing the score scale into a number of total score intervals of equal width. An example of this type of matching would be to form the matching variable by pooling pairs of adjacent total scores (e.g., 0 and 1, 2 and 3, etc.). This strategy may be seen as an attempt to increase reliability of assignment to levels of the matching variable.

The third strategy of forming the matching variable, total percent matching, is to allocate similar numbers of examinees (reference + focal) to each level of the matching variable. For example, five levels of the matching variable might be formed, corresponding to the quintiles of the pooled distribution.

The fourth strategy of the forming the matching variable, focal percent matching, is to pool cells such that each cell has a similar number of focal group members. An example of this type of matching would be to form five levels of the matching variable, corresponding to the quintiles of the focal group distribution.

The fifth major strategy of thick matching, censored matching, involves pooling extreme total scores (i.e., censoring the matching variable). Here extreme (highest and lowest) levels of the matching variable are pooled until each 2 X 2 cell at the extreme levels has some minimum number of observations (e.g., Zwick and Ercikan, 1989). When the test is easy (hard) for a group, total score matching may result in low (high) scores with no observations for that group. Levels are pooled until each cell of the most extreme levels contains some minimum number of observations.

The sixth major strategy, minimum frequency matching, is similar to censoring. This strategy pools adjacent cells until every level of the matching variable meets the minimum cell count criterion.

A seventh strategy was examined, in which no matching variable was used. This strategy measures impact, and was included to form an extreme with which to compare the other procedures for forming the matching variable.

In this study, we examined the various strategies proposed for thick matching, and compared the performance of these methods to the usual method of thin matching on total score. For some methods, more than one variation of the strategy was included. The strategies used are described below, and are summarized in Table 1.

Insert Table 1 about here

Method

This study used Monte Carlo methods to investigate the effects of various types of pooling on the MH procedure for detecting DIF. A primary challenge of the study was to select a manageable design which represented a reasonably diverse set of conditions. The levels of each factor were chosen to span a wide yet reasonable range with which test developers might routinely be faced.

Variables

Several independent variables were manipulated. The chief variables of interest were whether or not the studied item contained DIF, and the method of forming of the matching variable. In addition, four other independent variables were manipulated, due to their high salience and/or known effect in previous DIF research. One of these variables, the number of subjects, has not been routinely varied in previous MH DIF studies, so was of particular interest. The independent variables used were:

J - Number of criterion items, excluding the studied item, in the test (4 levels)--4, 9, 19, and 39 items.

S - Total number of subjects (reference group + focal group), (3 levels)--400, 800, 1600. In each case the

reference group comprised 75% of the sample, while the focal group was 25% of the sample.

M - Methods of forming the matching variable (11 levels) which are summarized in Table 1.

D - DIF in the studied item (2 levels)--no DIF, and DIF favoring the reference group.

A - IRT discrimination parameter **a** for the studied item (3 levels)--0.3, 1.0, 1.5.

B - Reference group IRT difficulty parameter $\mathbf{b}_R$ for the studied item (7 levels): -1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5. The focal group IRT difficulty parameter $\mathbf{b}_F$ differed with D.

Design of the Data Generation

In generating the data, the length of the criterion (J), and the number of subjects (S) were fully crossed to define 12 test conditions. DIF in the studied item (D), and difficulty (B) and discrimination (A) of the studied item were crossed within tests to define the studied items added to each test condition to form a test. Finally each mode of forming the matching variable (M) was applied to each test. Three specific combinations of method combined with test length were deleted from the study. These combinations are specified in Table 2 and are largely redundant with other cells in the table. Each replication of a test represented a test administration. Replications were made at two levels: replication within test ("administrations"), and replication of responses to the studied items for each administration. There were 20 replications for responses to the studied item and 20 replications of administrations within test conditions. Because of the size of the study, the replications of responses to the studied items for each administration were averaged for the analyses presented here.

Insert Table 2 about here

In design of analysis terms, averaging over the replications of studied item responses, the data were generated according to a split plot design, where variables defining the studied item (D, B, and A) and the method of matching (M) were within administration factors and variables determining the test condition (J and S) were between administration factors. The design for each criterion length (J) is illustrated in Table 3.

For each criterion length (J), each test condition is defined by the total number of subjects (S), listed along the left-hand side of the table. The variables defining the studied item are listed along the top of the table, along with the method of matching used in the analysis. The values that would appear in the body of the table for the data examined in this study are mean $\hat{\Delta}_{MH}$, se($\hat{\Delta}_{MH}$) or Mantel-Haenszel χ^2 values for 20 replications of responses to the studied item where the person parameters generated for the administration, the responses to the criterion items, and the variables defining the studied item were held constant.

Insert Table 3 about here

Data Generation Model and Procedure

Each administration (replication) of a test (i.e., combination of J, S, D, B and A) consisted of dichotomous data, with 0 reflecting an incorrect response, and 1 a correct response. No missing data were allowed. Simulated item responses were generated according to a three-parameter logistic (3PL) IRT model (Birnbaum, 1968). This model gives the probability of a correct response as:

$$P(\theta) = c + \frac{(1-c)}{1 + \exp[-1.7a(\theta-b)]} \qquad (7)$$

where:

θ is the proficiency which underlies the responses to the test items,

c is the guessing parameter, and corresponds to the probability that a subject of very low ability will get the item correct,

b the difficulty parameter, and

a is the discrimination parameter.

The discrimination parameter a describes the steepness of the ICC given in equation (7) at the point $\theta = b$. Large values of a indicate sharp discrimination between those subjects above and below b. The difficulty parameter b describes the point of maximum slope (discrimination) for an item, and corresponds to the proficiency for which the probability of a correct response is exactly half way between the value of the

guessing parameter and 1.0.

To generate the data for each test, NR proficiencies were sampled for the reference group, with $\theta_R \sim N(0,1)$, and NF proficiencies were sampled for the focal group, with $\theta_F \sim N(-1,1)$. Next, IRT item parameters were then sampled for each item in the criterion. Item discrimination parameters were sampled from a log-normal distribution, where $a = \exp(z)$, with $z \sim N(0,0.1225)$. Therefore, the mean a-value was 1.0. Difficulty parameters for the criterion were sampled from N(0,1). These distributions were selected to approximate the empirical SAT item parameters given in Lord (1968). A common c-value was used for all items, with $c \equiv 0.2$.

DIF in the studied item, discrimination of the studied item and difficulty of the studied item were crossed within tests, to define 42 studied items for each test. Twenty-one of these items displayed DIF, while the other 21 had no DIF. For the reference group, the b parameter was defined by the difficulty of the studied item, as described above. For items without DIF, $b_F = b_R$, while for items with DIF $b_F = b_R + 0.3$. Twenty replicates were generated for each studied item, yielding a total of 840 studied items per test.

Finally, each subject's probability of getting each item correct P_{ij} was computed according to equation (7). A uniform [0,1] pseudo-random number v_{ij} was then generated. The item was considered correct if $v_{ij} \leq P_{ij}$, and incorrect otherwise.

Uniform pseudo-random numbers were generated using the algorithm RANDOM (Wichman and Hill, 1983). Normal deviates were generated by applying a probit transformation to a uniform random number using algorithm PPDN (Beasley and Springer, 1977). Data generation and MH analyses were conducted using FORTRAN-77 programs written by the authors. The programs were compiled using Microsoft's FORTRAN compiler, version 5.0. All programs were run on a COMPAQ Deskpro 386/20e, with a 80387 math coprocessor.

DIF Analyses

The dichotomous matrix of item responses was then analyzed using the MH procedure with each method of forming the matching variable. MH analyses were always conducted by including the studied item in the matching variable.

For analysis, a test was composed of the criterion items and the studied item. So, the number of items in a test, and contributing to the matching variable, was 5, 10, 20 or 40. Each method of forming the matching variable was applied to each replicate of each studied item of each administration of each test. For each analysis, the K levels of the matching variable were defined according to the appropriate strategy for pooling data.

Three measures of DIF were recorded for each analysis, the DIF measure $\hat{\Delta}_{MH}$, its standard error, and the MH χ^2. The mean and standard deviation of the 20 replicates of each studied item was computed for each measure. These means and standard deviations were then output, and served as the dependent variables for the study.

Finally, this entire process (sampling item parameters and subject proficiencies, generating item responses and performing MH analyses) was repeated 20 times for each of the 12 test conditions. Mean MH measures of DIF ($\hat{\Delta}_{MH}$, se($\hat{\Delta}_{MH}$) and χ^2) were then analyzed to determine the effects of the various manipulations on the DIF measures obtained.

As described above, there were up to 27,720 (3 X 2 X 7 X 3 X 11 X 20) observations generated for each of the four criterion lengths, and they were generated in a complex way. Because the purpose of the study was to compare each method of matching to traditional thin matching and because amount of data that was generated was large, results are illustrated by appropriate means for specific test and studied item conditions

Results

The primary independent variable of interest was the method of forming the matching variable. Two criteria will be used to assess the effectiveness of the methods:

1) Does the thick matching method result in substantial confounding of DIF with impact? This question will be examined by comparing the means of the methods when D=0. For $\hat{\Delta}_{MH}$, the expected mean for this condition is 0.0, while for χ^2 the expected mean is 1.0. The mean for thin matching may also be used as a guide.

2) Does the method yield a large difference between items which exhibit DIF and those which do not? Is the difference larger than that obtained using thin matching?

The primary information needed to evaluate these criteria are means for method, and its interaction with the presence or absence of DIF in the studied item. The means of this interaction for $\hat{\Delta}_{MH}$ are presented in Table 4.

Insert Table 4 about here

$\hat{\Delta}_{MH}$ Measure

Three of the methods, Censor (20), Min. Freq. (20), and No Matching, produce means which are substantially different from 0.0 when there was no DIF in the studied item (D=0). When D=0, the mean $\hat{\Delta}_{MH}$ for each of these methods is <u>larger</u> than when D=1 for the thin matching. Therefore, we conclude that these three methods demonstrate substantial confounding of DIF and impact, and hence they will not be discussed further.

Two other methods, Censor (5) and Min. Freq. (5) show some evidence of confounding DIF with impact. The $\hat{\Delta}_{MH}$ values for the D=0 condition are twice as large as those for any of the six remaining methods. The difference between the D=0 and D=1 conditions is also less than that of any of the remaining methods. Thus, there is some evidence that these two methods may not be satisfactory, but it is

not conclusive.

The relative sparseness of the MH 2 X 2 X K table is primarily determined by the specific combination of the number of items (J), the number of subjects (S) and the method of forming the matching variable. Table 5 gives the mean $\hat{\Delta}_{MH}$ for each of the J*S*M combinations, when DIF=0.

Insert Table 5 about here

Table 5 provides further evidence that Censor (5) and Min. Freq. (5) strategies perform poorly. These methods have negative means (suggesting appreciable confounding of DIF with impact) larger in magnitude than those for the other methods, particularly for small values of S, i.e., S=400. Because one is most interested in using thick matching when the 2 X 2 X K table is sparse, this property makes these methods particularly unattractive. They will not be discussed further.

Table 5 also reveals that thin matching performs poorly for very short tests. When J=4 and J=9, the $\hat{\Delta}_{MH}$ value is substantially positive for all sample sizes, falsely indicating DIF against the Reference group. When J=9 this effect is much smaller when S=800 and S=1600. This suggests that, for short tests, the $\hat{\Delta}_{MH}$ statistic may perform satisfactorily for larger sample sizes, although it would be necessary to extend this study to more levels of numbers of subjects than was examined in this study.

The Percent Total and Percent Focal methods show an opposite pattern of results to those of Censor (5) and Min. Freq. (5). These two methods have a tendency to produce negative means for larger numbers of subjects. The strength of this effect is increased as the length of the test increases. For J=19 and J=39, both methods show substantial confounding of DIF with impact for all sample sizes. For these test lengths, thin matching is clearly superior to both methods. For J=9, all three methods (Thin Matching, Percent Total, and Percent Focal matching methods) have drawbacks, and which is preferred depends upon the number of subjects. Finally, for all conditions, Percent Total performed slightly better than did Percent focal.

Of the three remaining methods, Equal Interval, Censor (1) and Min. Freq. (1), all yield $\hat{\Delta}_{MH}$ means

which are reasonably close to the expected value of 0.0 when D=0. For test lengths of J=19 or less, these three methods uniformly yield mean $\hat{\Delta}_{MH}$ values which are closer in absolute value to 0.0 than the mean for thin matching.

Table 6 gives mean differences between the D=0 and D=1 cases, along with mean standard errors for the four methods of special interest, Thin Matching, Equal Interval, Censor (1) and Min. Freq. (1). The standard errors given in Table 6 are only for the D=1 condition, because those observed for D=0 were all within .01 of those for D=1.

Insert Table 6 about here

Table 6 reveals that the four methods are very similar in their mean differences (D=1 - D=0) and mean standard errors. Both the effect size and the standard error are slightly larger for Thin Matching, than for Equal Interval method, which is in turn slightly larger than for Censor(1) and Min. Freq. (1). As both the effect size and standard error are ordered in the same direction, and differences are small, their effects will tend to balance. Given this, there is clear evidence that the Min. Freq. (1) and Censor (1) methods of thick matching will yield better results in DIF analyses which must be conducted on small samples. The Equal Interval method also has promise in improving the performance of DIF analyses beyond that of thin matching.

For Equal Interval, Censor(1) and Min. Freq. (1) matching methods, the separation of values for DIF and non-DIF items is similar to that for thin matching; the interval between the means for non-DIF and DIF items is similar for each method. The improvement comes in the more appropriate location of the interval for the thick matching procedures. The difference in the location of the interval would be expected to have the effect that DIF items are more likely to be identified as such, and fewer non-DIF items will be falsely identified as exhibiting DIF against the reference group. As both of these outcomes are highly desirable, the use of a mild thick matching method such as Censor (1), Min. Freq. (1), or the Equal Interval method used here is recommended.

χ^2 Measure

As was the case for the $\hat{\Delta}_{MH}$ measure, the chief measure needed to evaluate the criteria listed above are means for each method, and the interaction of the method with the presence or absence of DIF in the studied item. The means of this interaction for χ^2 are presented in Table 7.

Insert Table 7 about here

As was the case for $\hat{\Delta}_{MH}$, the results for Censor (20), Min. Freq. (20), and No Matching show a large confounding of DIF with impact. Hence, these methods are discarded from further examination. Also, as was the case for $\hat{\Delta}_{MH}$, the mean χ^2 for Censor (5) and Min. Freq. (5) appear somewhat larger than those for other conditions, but this effect is not overwhelming based on the results in Table 7.

To further examine the χ^2, the means for the J*S*M interaction (when D=0) are given in Table 8. The expected value for the table entries is 1.0.

Insert Table 8 about here

Table 8 reveals two interesting results about the behavior of the Thin Matching method. First, for every test length, the value of the χ^2 increases as the number of subjects increases; the mean for J=4, S=1600 is almost significant at $\alpha=.05$. Second, this sample size effect decreases as the test gets longer.

The mean χ^2 values for methods Censor (5) and Min. Freq. (5) are uniformly larger than those for Thin Matching. While this is not a large effect, it is consistent; only J=4, S=1600 gives an opposite result. Thus, as in the case of $\hat{\Delta}_{MH}$ values, these methods generally yield results which are slightly more likely to falsely identify items as containing DIF. Thus, these methods do not present any advantage over thin matching, and will not be discussed further.

On the other hand, the mean χ^2 values for the other thick matching methods are similar to those for thin matching when D=0, although most are slightly smaller than those for thin matching. This includes the methods Censor (1) and Min. Freq. (1). While this is not a large effect, it is indication that one or more of

these methods may present some advantage over thin matching.

To better compare the methods, the mean χ^2 values for the D=1 condition were computed, and the cell mean χ^2 values for the D=0 condition were subtracted from them. The result is the expected difference in χ^2 between items which demonstrate DIF and those which do not. Differences larger than those for thin matching are noteworthy. These results are summarized in Table 9.

Insert Table 9 about here

The results in Table 9 reveal that all of the five remaining thick matching methods provide an advantage over thin matching in detecting DIF when it is present. For Censor (1) and Min. Freq. (1), the advantage is relatively small. The Equal Interval method produces a somewhat larger improvement, but the Percent Total and Percent Focal methods consistently outperform all of the other methods, with Percent Focal yielding the largest gains over thin matching.

The remarkable thing is that the Percent Focal method has a mean χ^2 which is similar to or somewhat less than that for thin matching when D=0. Yet, when D=1, the χ^2 value for the Percent Focal method is approximately 25% larger than that for thin matching. It seems that, when the Mantel-Haenszel χ^2 is to be used as the measure of DIF, some version of the Percent Focal method of thick matching can yield real improvements in detecting DIF.

Other Factors Affecting DIF

Two other factors, discrimination of the studied item (A) and difficulty of the studied item (B) were also varied in this study. These two factors were not of primary interest for the purposes of this paper. Rather, they were included to insure that their effects were systematically present, and so would not confound the above results concerning methods of forming the matching variable. Still, it is appropriate here to briefly consider the most salient aspects of these two studied item variables, and their interaction with DIF.

Examining the pattern of results revealed that the studied item variables A and B have large main effects and large two-way and three-way interactions with each other and with D, although the size of the effects is to some extent due to the number of levels chosen and their values.

Table 10 gives cell $\hat{\Delta}_{MH}$ means (using thin matching) for the three-way A*B*D interaction. These results are for the condition J=39 and S=1600, so thin matching should function well and the effects are not artifacts of the method of forming the matching variable.

Insert Table 10 about here

The mean $\hat{\Delta}_{MH}$ value increases with the level of B. For very easy items, the mean $\hat{\Delta}_{MH}$ is substantially negative, while for hard items the value is slightly positive. Increasing B also decreases the difference between the means for non-DIF and DIF items. Increasing the discrimination A makes the B differences larger, especially for easy items. Increasing A also increases the difference between the means for non-DIF and DIF items.

These findings closely parallel those of Donoghue, Holland and Thayer (1989), and Allen and Donoghue (1991), and the result for item discrimination also parallels a finding in Camilli and Smith (1990). A future paper will fully explore the interactions of studied item characteristics with the characteristics of the criterion.

Summary and Conclusions

This paper sought to examine whether "thick matching" (pooling levels of the Mantel-Haenszel matching variable) could improve the performance of the MH procedure for detecting DIF. The answer clearly is yes. For short tests (5 or 10 items), thin matching yielded very poor results, with a tendency to falsely identify items as possessing DIF against the reference group. This finding supports the suggestion in Donoghue, Holland and Thayer (1989) that the usual MH procedure should not be used for such short tests. For long tests (40 items), especially with adequate sample sizes (1600), thin matching yielded the best results

of any method examined. Intermediate length tests yielded similar results for thin matching and the best methods of thick matching.

The method of thick matching which performs best depends upon the measure used to detect DIF. When the outcome measure is $\hat{\Delta}_{MH}$, a mild form of thick matching is best. Three methods were identified as having promise. These methods were: (a) Censoring the matching variable (combining extreme levels until each cell of the extreme 2 X 2 tables has at least one observation); (b) Minimum Frequency matching (pooling levels until each cell of the 2 X 2 table has at least one observation every level of the matching variable); and (c) Equal Interval matching, in which every 2 levels of the matching variable are combined (scores 0 and 1 are combined, 2 and 3 are combined, etc.).

The Censoring and Min. Freq. methods would be expected to yield similar or identical results unless the 2 X 2 X K table is very sparse. From the levels of variables studied here, it is not possible to meaningfully separate these two methods. However, it seems that for shorter tests (20 items or fewer), either of these strategies will yield improvement over usual, thin matching. Thus, they may prove helpful for short tests and performing DIF analyses on subscales.

When the χ^2 measure was used to detect DIF, a thick matching based on pooling to yield approximately equal numbers of examinees (Percent Total) or equal numbers of focal group members (Percent Focal) yielded the best results. Percent Focal proved slightly superior to Percent Total, but the differences were small. Both were superior to thin matching for shorter tests, yielding a χ^2 closer to the expected value for non-DIF items, and a larger χ^2 value than thin matching when the studied item possessed DIF.

The Equal Interval method for $\hat{\Delta}_{MH}$ and the Percent Focal and Percent Total methods for χ^2 should be studied further. Each method was represented by one, somewhat arbitrary version of the method. More systematic study of several variations of these methods, as well as variations (such as Scheuneman's 1979 suggestion that levels of the matching variable have equal numbers of Focal group members who pass the item) is warranted.

The effect of studied item characteristics was also briefly examined. Both difficulty and discrimination of the studied item were found to have a strong effect on the value of $\hat{\Delta}_{MH}$. It may be that DIF analyses may be made more precise by taking into account characteristics of the studied item.

While the approach taken here has been strictly in terms of creating intervals of total scores to use as matching variables, the conceptual issues are related to those of the use of auxiliary information, (as in Zwick and Ercikan, 1989). Using auxiliary information in addition to the total score may be seen as matching to a finer degree than matching done on total score alone. Thus, information from this study about the relationship between the thickness of the matching and the validity of the MH procedure also has implications for this related type of matching.

Finally, one important limitation of this study is emphasis on descriptive statistics. While the means of various measures are good indicators of directions of effects, there is no guarantee that effects of the size observed will yield many differences in the decisions made about individual items. Having here identified plausible methods for thick matching, an important next step is to simulate the decision making that the various thick matching strategies would yield. This will allow us to determine whether these effects are large enough to be worth the added bother and expense of accounting for them.

References

Allen, N. L. & Donoghue, J. R. (1991). Applying the Mantel-Haenszel procedure to complex sample data. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL, April 1991.

Beasley, J. D. & Springer, S. G. (1977). Algorithm AS 111: The percentage points of the normal distribution. Applied Statistics, 26, 118-121.

Breslow, N. (1981). Odds ratio estimates when the data are sparse. Biometrika, 68, 73-84.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, Statistical theories of mental test scores, 395-479. Menlo Park, CA: Addison-Wesley Publishing Company.

Camilli, G. (1979). A critique of the chi-square method for assessing item bias. Unpublished paper, Laboratory of Educational Research, University of Colorado.

Camilli, G. & Smith, J. K. (1990). Comparison of the Mantel-Haenszel test with a randomized and a jackknife test for detecting biased items. Journal of Educational Statistics, 15, 53-67.

Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rheinhart, and Winston.

Donner, A. & Hauck, W. W. (1986) The large sample relative efficiency of the Mantel-Haenszel estimator in the fixed strata case. Biometrics, 42, 537-545.

Donoghue, J. R., Holland, P. W. & Thayer, D.T. (1989). A Monte Carlo study of factors that affect the Mantel-Haenszel measure of differential item functioning. Paper presented at the NCME symposium on Methodological Issues Concerning the Study of Differential Item Functioning. San Francisco, March, 1989.

Holland, P. W. (1985). On the study of differential item performance without IRT. Proceedings of the Military Testing Association, October, 1985.

Holland, P. W. & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel Procedure. In H. Wainer and H. Braun (Eds.), Test validity. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M. (1968). An analysis of the verbal Scholastic Achievement Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 28, 989-1020.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Phillips, A. & Holland, P.W. (1987). Estimation of the variance of the Mantel-Haenszel log-odds-ratio estimate. Biometrika, 43, 425-431.

Scheuneman, J. (1979). A method of assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Scheuneman, J. D. (1981). A response to Baker's criticism. Journal of Educational Measurement, 18, 63-66.

Thissen, D., Steinberg, L. & Wainer, H. (1988). Using item response theory in the study of group difference in tracelines. In H. Wainer and H. Braun (Eds.), Test validity. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wichman, B. A. & Hill, I. D. (1982). Algorithm AS 183: An efficient and portable pseudo-random number generator. Applied Statistics, 31, 188-190.

Zwick, R. (1990). When item response function and Mantel-Haenszel definitions of differential item functioning coincide? Journal of Educational Statistics, 15, 185-197.

Zwick, R. & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP History Assessment. Journal of Educational Measurement, 26, 55-66.

Table 1

Descriptions of Strategies Used to Form Matching Variable

Method	Abbreviation	Description
Thin Matching Variable	Thin Matching	Levels of table defined by total score on test
Equal Interval Matching Variable	Equal Interval	Pairs of total scores are combined to define levels of the table (i.e., 0 and 1 are combined, 2 and 3 are combined, etc.).
Percent of Total Sample Matching Variable	Percent Total	Score Intervals are combined to approximate quintiles of the combined (reference + focal) sample.
Percent of Focal Sample Matching Variable	Percent Focal	Score Intervals are combined to approximate quintiles of the focal group sample.
Censored Matching Variable	Censor (1)	Extreme (high and low) score intervals are pooled until the 2 X 2 tables for the lowest and the highest levels each have at least 1 observation. Only extreme levels are pooled.
"	Censor (5)	Extreme (high and low) score intervals are pooled until the 2 X 2 tables for the lowest and the highest levels each have at least 5 observations. Only extreme levels are pooled.
"	Censor (20)	Extreme (high and low) score intervals are pooled until the 2 X 2 tables for the lowest and the highest levels each have at least 20 observations. Only extreme levels are pooled.
Minimum Cell Frequency Matching Variable	Min. Freq. (1)	Score intervals are pooled until the 2 X 2 tables for every level has at least 1 observation.
"	Min. Freq. (5)	Score intervals are pooled until the 2 X 2 tables for every level has at least 5 observations.
"	Min. Freq. (20)	Score intervals are pooled until the 2 X 2 tables for every level has at least 20 observations.
No Matching Variable	No Matching	No matching is performed. All observations are pooled into a single 2 X 2 table.

BEST COPY AVAILABLE

Table 2

Design Specifications for
Methods of Forming Matching Variable
and Length of Criterion

Mode of Forming the Matching Variable	Length of Criterion			
	4	9	19	39
Thin Matching				
Equal Interval	No			
Percent Total	No			
Percent Focal	No			
Censor (1)				
Censor (5)				
Censor (20)				
Min. Freq. (1)				
Min. Freq. (5)				
Min. Freq. (20)				
No Matching				

Table 3
Design of the Data Generation for each Criterion Length

S	DATA SET	DIF							NO DIF							DIF
		-1.5			...	1.5			-1.5			...	1.5			B
		0.3	1.0	1.5	...	0.3	1.0	1.5	0.3	1.0	1.5	...	0.3	1.0	1.5	A
		1..11	1..11	1..11	...	1..11	1..11	1..11	1..11	1..11	1..11	...	1..11	1..11	1..11	M
400	1 2 3 . . . 20															
800	1 2 3 . . . 20															
1600	1 2 3 . . . 20															

Table 4

Mean $\hat{\Delta}_{MH}$ for Methods of Forming the Matching Variable

METHOD	D=0	D=1	Difference (D=0 - D=1)
Thin Matching	0.07694	-0.42378	0.50072
Equal Interval	-0.03617	-0.53139	0.49522
Percent Total	-0.05744	-0.54097	0.48353
Percent Focal	-0.09589	-0.59460	0.49861
Censor (1)	-0.01050	-0.50196	0.49146
Censor (5)	-0.18222	-0.64565	0.46343
Censor (20)	-0.92321	-1.33334	0.41013
Min. Freq. (1)	-0.01381	-0.50391	0.49010
Min. Freq. (5)	-0.19545	-0.65911	0.46366
Min. Freq. (20)	-0.94017	-1.35183	0.41165
No Matching	-1.63352	-2.03623	0.40270

Table 5

Mean $\hat{\Delta}_{MH}$ Values for Methods by Length of Criterion (J) and Number of Subjects (S), DIF=0

S	METHOD	Length of Criterion			
		4 Items	9 Items	19 Items	39 Items
400	Thin Matching	0.14654	0.14133	0.06215	0.00384
	Equal Interval	*	0.04828	0.03142	-0.01196
	Percent Total	*	-0.00233	-0.08453	-0.17179
	Percent Focal	*	-0.07034	-0.14107	-0.18600
	Censor (1)	-0.06862	0.00285	-0.01780	-0.06558
	Censor (5)	-0.49490	-0.32107	-0.28018	-0.30565
	Min. Freq. (1)	-0.06926	-0.00413	-0.02534	-0.08028
	Min. Freq. (5)	-0.49892	-0.35657	-0.32030	-0.33111
800	Thin Matching	0.20325	0.06226	0.01365	0.01307
	Equal Interval	*	-0.02797	-0.01147	0.01187
	Percent Total	*	-0.06547	-0.11815	-0.13484
	Percent Focal	*	-0.11046	-0.16999	-0.18378
	Censor (1)	-0.01882	0.00736	-0.01718	-0.02158
	Censor (5)	-0.20337	-0.13799	-0.12789	-0.12184
	Min. Freq. (1)	-0.01882	0.00592	-0.02002	-0.02241
	Min. Freq. (5)	-0.20363	-0.14519	-0.14128	-0.14073
1600	Thin Matching	0.14922	0.05453	0.04791	0.02552
	Equal Interval	*	-0.04707	0.01245	0.01208
	Percent Total	*	-0.08216	-0.09865	-0.11264
	Percent Focal	*	-0.13747	-0.15465	-0.19506
	Censor (1)	0.01346	0.02027	0.02819	0.01149
	Censor (5)	-0.05934	-0.06180	-0.02884	-0.04375
	Min. Freq. (1)	0.01346	0.01983	0.02678	0.00860
	Min. Freq. (5)	-0.05934	-0.06285	-0.03343	-0.05210

* This cell is empty by design.

Table 6

Mean Differences between $\hat{\Delta}_{MH}$ Values for DIF=1 and DIF=0
for Methods by Length of Criterion (J) and Number of Subjects (S)
(D=1 - D=0)

S	METHOD	Length of Criterion			
		4 Items	9 Items	19 Items	39 Items
400	Thin Matching	0.47306	0.47922	0.52069	0.54292
	Equal Interval	*	0.47670	0.51330	0.53909
	Censor (1)	0.45510	0.46183	0.50341	0.52923
	Min. Freq. (1)	0.45472	0.46076	0.50080	0.52175
800	Thin Matching	0.47517	0.49332	0.51243	0.52883
	Equal Interval	*	0.48470	0.51004	0.52795
	Censor (1)	0.46936	0.48623	0.50552	0.52072
	Min. Freq. (1)	0.46941	0.48562	0.50479	0.51835
1600	Thin Matching	0.47929	0.48307	0.50271	0.51790
	Equal Interval	*	0.47661	0.50015	0.51674
	Censor (1)	0.47461	0.47885	0.49876	0.51394
	Min. Freq. (1)	0.47461	0.47863	0.49831	0.51348

(continued on following page)

Table 6 (continued)

Mean SE($\hat{\Delta}_{MH}$) Values for Methods by Length of Criterion (J) and Number of Subjects (S), DIF=1

S	METHOD	Length of Criterion			
		4 Items	9 Items	19 Items	39 Items
400	Thin Matching	0.7319	0.7020	0.6985	0.7061
	Equal Interval	*	0.6930	0.6879	0.6928
	Censor (1)	0.7136	0.6966	0.6933	0.6993
	Min. Freq. (1)	0.7135	0.6959	0.6904	0.6894
800	Thin Matching	0.5165	0.4897	0.4823	0.4871
	Equal Interval	*	0.4843	0.4776	0.4815
	Censor (1)	0.5061	0.4883	0.4808	0.4852
	Min. Freq. (1)	0.5061	0.4883	0.4804	0.4830
1600	Thin Matching	0.3647	0.3440	0.3410	0.3401
	Equal Interval	*	0.3401	0.3389	0.3379
	Censor (1)	0.3592	0.3436	0.3406	0.3396
	Min. Freq. (1)	0.3592	0.3435	0.3406	0.3393

*This cell is empty by design.

Table 7

Mean χ^2 for Methods of Forming the Matching Variable

METHOD	D=0	D=1	Difference (D=1 - D=0)
Thin Matching	1.8813	3.9177	2.0365
Equal Interval	1.8694	4.4653	2.5959
Percent Total	1.9974	4.7091	2.7117
Percent Focal	1.8545	4.7085	2.8540
Censor (1)	1.7975	4.1060	2.3085
Censor (5)	2.1003	4.9115	2.8112
Censor (20)	7.6728	13.3595	5.6867
Min. Freq. (1)	1.8001	4.1215	2.3214
Min. Freq. (5)	2.1259	5.0010	2.8751
Min. Freq. (20)	7.7499	13.5817	5.8318
No Matching	22.1190	34.9683	12.8493

Table 8

Mean χ^2 Values for Methods by Length of Criterion (J) and Number of Subjects (S), DIF=0

S	METHOD	Length of Criterion			
		4 Items	9 Items	19 Items	39 Items
400	Thin Matching	1.80119	1.29710	0.99372	0.83527
	Equal Interval	*	1.28807	0.99729	0.83535
	Percent Total	*	1.28780	1.06274	0.95464
	Percent Focal	*	1.18927	0.98832	0.92707
	Censor (1)	1.75072	1.25970	0.99140	0.84433
	Censor (5)	2.83066	1.66227	1.28420	1.08702
	Min. Freq. (1)	1.75478	1.26455	0.99080	0.84439
	Min. Freq. (5)	2.83510	1.72439	1.35840	1.16527
800	Thin Matching	2.65112	1.87036	1.22318	1.00886
	Equal Interval	*	1.89491	1.23101	1.02130
	Percent Total	*	1.96267	1.33428	1.15013
	Percent Focal	*	1.74931	1.28051	1.13497
	Censor (1)	2.29773	1.84355	1.23582	1.01121
	Censor (5)	2.70990	2.06788	1.39917	1.13921
	Min. Freq. (1)	2.29773	1.84955	1.23999	1.01699
	Min. Freq. (5)	2.71185	2.07651	1.42054	1.17519
1600	Thin Matching	4.75205	3.08913	1.78448	1.26867
	Equal Interval	*	3.19077	1.81649	1.28378
	Percent Total	*	3.24768	2.00324	1.50318
	Percent Focal	*	2.82155	1.80317	1.51377
	Censor (1)	4.22249	3.05090	1.78237	1.27935
	Censor (5)	4.60978	3.15707	1.89946	1.35697
	Min. Freq. (1)	4.22249	3.05369	1.78525	1.28155
	Min. Freq. (5)	4.60978	3.16062	1.90795	1.36569

* This cell is empty by design.

BEST COPY AVAILABLE

Table 9

Mean Differences (D=1 - D=0) for χ^2

S	Method	4 Items	9 Items	19 Items	39 Items
400	Thin Matching	0.80926	0.74157	0.85206	0.84632
	Equal Interval	*	0.95261	0.93723	0.90109
	Percent Total	*	1.01004	1.17398	1.25725
	Percent Focal	*	1.09372	1.25728	1.28727
	Censor (1)	1.18662	0.98493	0.99151	0.97034
	Min. Freq. (1)	1.18479	0.99551	1.01659	1.01450
800	Thin Matching	1.35571	1.91488	1.89888	1.77211
	Equal Interval	*	2.30035	2.04783	1.82391
	Percent Total	*	2.42231	2.53794	2.50244
	Percent Focal	*	2.54812	2.72411	2.69781
	Censor (1)	2.06080	2.11314	2.01040	1.90443
	Min. Freq. (1)	2.06154	2.11575	2.02419	1.92022
1600	Thin Matching	3.18524	3.97182	3.67270	3.41715
	Equal Interval	*	4.89443	4.03272	3.58073
	Percent Total	*	5.09546	4.96557	4.69758
	Percent Focal	*	5.36888	5.34994	5.39779
	Censor (1)	3.97882	4.18615	3.79803	3.51683
	Min. Freq. (1)	3.97882	4.18880	3.81063	3.54519

* This cell is empty by design.

Table 10

Mean Values of $\hat{\Delta}_{MH}$ for A*B*D Interaction

(J=39, S=1600)

A	B	D=0	D=1	Difference (D=0 - D=1)
0.3	-1.5	0.11456	-0.16417	0.27873
	-1.0	0.15143	-0.16819	0.31961
	-0.5	0.16513	-0.14980	0.31493
	0.0	0.17722	-0.10494	0.28216
	0.5	0.19618	-0.09439	0.29056
	1.0	0.18936	-0.04065	0.23000
	1.5	0.18432	-0.01565	0.19996
1.0	-1.5	-0.29788	-1.23295	0.93507
	-1.0	-0.20292	-1.02012	0.81720
	-0.5	-0.06922	-0.83253	0.76331
	0.0	-0.00114	-0.60259	0.60145
	0.5	0.11095	-0.36394	0.47488
	1.0	0.19466	-0.15178	0.34644
	1.5	0.22063	0 02728	0.19334
1.5	-1.5	-0.61760	-1.81095	1.19335
	-1.0	-0.36036	-1.46414	1.10378
	-0.5	-0.16927	-1.07428	0.90501
	0.0	-0.00920	-0.74172	0.73253
	0.5	0.07406	-0.38637	0.46043
	1.0	0.21522	-0.06405	0.27927
	1.5	0.26974	0.11593	0.15381